

Online Learning

Lecture 3

Nicolò Cesa-Bianchi

Università degli Studi di Milano

FTRL recap

FTRL:

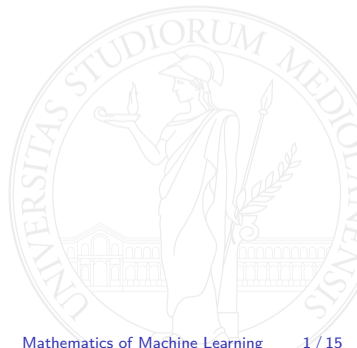
$$\mathbf{w}_{t+1} = \operatorname{argmin}_{\mathbf{w} \in \mathbb{V}} \psi(\mathbf{w}) + \sum_{s=1}^t \ell_s(\mathbf{w}_s)$$

Best model in \mathbb{V}

$$\mathbf{w}^* = \operatorname{argmin}_{\mathbf{w} \in \mathbb{V}} \sum_{t=1}^T \ell_t(\mathbf{w})$$

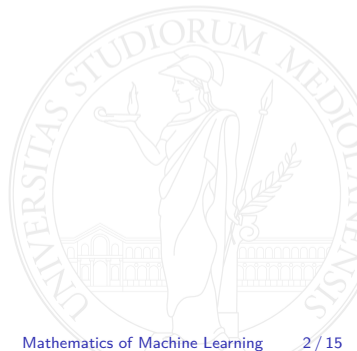
Regret

$$R_T = \sum_{t=1}^T \ell_t(\mathbf{w}_t) - \sum_{t=1}^T \ell_t(\mathbf{w}^*)$$



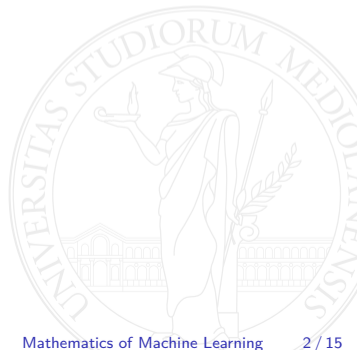
FTRL = FTL + regularization

► $L_t = \ell_0 + \ell_1 + \dots + \ell_t$ where $\ell_0 = \psi$



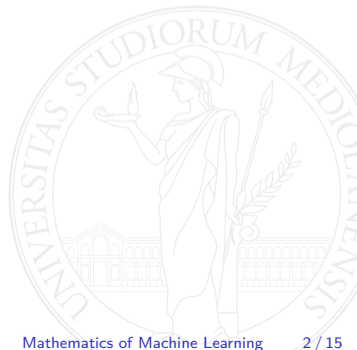
FTRL = FTL + regularization

- ▶ $L_t = \ell_0 + \ell_1 + \dots + \ell_t$ where $\ell_0 = \psi$
- ▶ $\mathbf{w}_1 = \operatorname{argmin}_{\mathbf{w} \in \mathcal{V}} \ell_0(\mathbf{w}) = \operatorname{argmin}_{\mathbf{w} \in \mathcal{V}} \psi(\mathbf{w})$



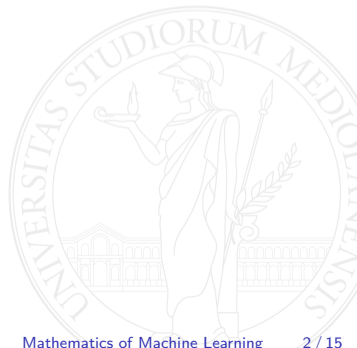
FTRL = FTL + regularization

- ▶ $L_t = \ell_0 + \ell_1 + \dots + \ell_t$ where $\ell_0 = \psi$
- ▶ $\mathbf{w}_1 = \operatorname{argmin}_{\mathbf{w} \in \mathcal{V}} \ell_0(\mathbf{w}) = \operatorname{argmin}_{\mathbf{w} \in \mathcal{V}} \psi(\mathbf{w})$
- ▶ $\mathbf{w}_{t+1} = \operatorname{argmin}_{\mathbf{w} \in \mathcal{V}} L_t(\mathbf{w})$



FTRL = FTL + regularization

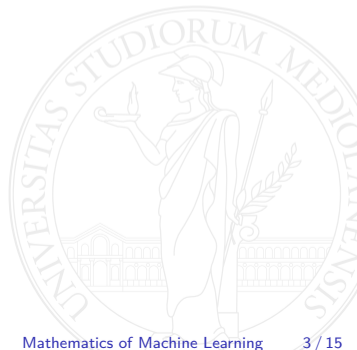
- ▶ $L_t = \ell_0 + \ell_1 + \dots + \ell_t$ where $\ell_0 = \psi$
- ▶ $\mathbf{w}_1 = \operatorname{argmin}_{\mathbf{w} \in \mathcal{V}} \ell_0(\mathbf{w}) = \operatorname{argmin}_{\mathbf{w} \in \mathcal{V}} \psi(\mathbf{w})$
- ▶ $\mathbf{w}_{t+1} = \operatorname{argmin}_{\mathbf{w} \in \mathcal{V}} L_t(\mathbf{w})$
- ▶ We give a different proof of the FTL stability lemma



FTRL stability lemma

$$\sum_{t=0}^T \ell_t(\mathbf{w}_{t+1}) \leq \inf_{\mathbf{u} \in \mathbb{V}} \sum_{t=0}^T \ell_t(\mathbf{u})$$

(we prove this by induction on T)

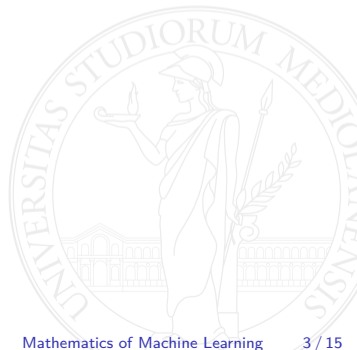


FTRL stability lemma

$$\sum_{t=0}^T \ell_t(\mathbf{w}_{t+1}) \leq \inf_{\mathbf{u} \in \mathbb{V}} \sum_{t=0}^T \ell_t(\mathbf{u})$$
$$\ell_0(\mathbf{w}_1) \leq \inf_{\mathbf{u} \in \mathbb{V}} \ell_0(\mathbf{u})$$

(we prove this by induction on T)

(base case $T = 0$: $\mathbf{w}_1 = \operatorname{argmin}_{\mathbf{w} \in \mathbb{V}} \ell_0(\mathbf{w})$)



FTRL stability lemma

$$\sum_{t=0}^T \ell_t(\mathbf{w}_{t+1}) \leq \inf_{\mathbf{u} \in \mathbb{V}} \sum_{t=0}^T \ell_t(\mathbf{u})$$

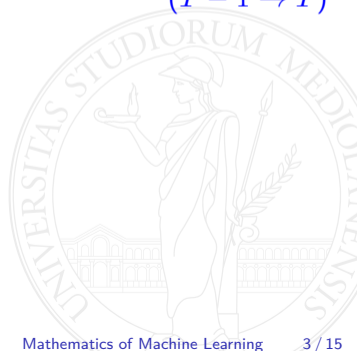
$$\ell_0(\mathbf{w}_1) \leq \inf_{\mathbf{u} \in \mathbb{V}} \ell_0(\mathbf{u})$$

$$\sum_{t=0}^{T-1} \ell_t(\mathbf{w}_{t+1}) \leq \inf_{\mathbf{u} \in \mathbb{V}} \sum_{t=0}^{T-1} \ell_t(\mathbf{u})$$

(we prove this by induction on T)

(base case $T = 0$: $\mathbf{w}_1 = \operatorname{argmin}_{\mathbf{w} \in \mathbb{V}} \ell_0(\mathbf{w})$)

($T - 1 \rightarrow T$)



FTRL stability lemma

$$\sum_{t=0}^T \ell_t(\mathbf{w}_{t+1}) \leq \inf_{\mathbf{u} \in \mathbb{V}} \sum_{t=0}^T \ell_t(\mathbf{u})$$

$$\ell_0(\mathbf{w}_1) \leq \inf_{\mathbf{u} \in \mathbb{V}} \ell_0(\mathbf{u})$$

$$\sum_{t=0}^{T-1} \ell_t(\mathbf{w}_{t+1}) \leq \inf_{\mathbf{u} \in \mathbb{V}} \sum_{t=0}^{T-1} \ell_t(\mathbf{u})$$

$$\sum_{t=0}^{T-1} \ell_t(\mathbf{w}_{t+1}) \leq \sum_{t=0}^{T-1} \ell_t(\mathbf{w}_{T+1})$$

(we prove this by induction on T)

(base case $T = 0$: $\mathbf{w}_1 = \operatorname{argmin}_{\mathbf{w} \in \mathbb{V}} \ell_0(\mathbf{w})$)

($T - 1 \rightarrow T$)

(choose $\mathbf{u} = \mathbf{w}_{T+1}$)

FTRL stability lemma

$$\sum_{t=0}^T \ell_t(\mathbf{w}_{t+1}) \leq \inf_{\mathbf{u} \in \mathbb{V}} \sum_{t=0}^T \ell_t(\mathbf{u})$$

$$\ell_0(\mathbf{w}_1) \leq \inf_{\mathbf{u} \in \mathbb{V}} \ell_0(\mathbf{u})$$

$$\sum_{t=0}^{T-1} \ell_t(\mathbf{w}_{t+1}) \leq \inf_{\mathbf{u} \in \mathbb{V}} \sum_{t=0}^{T-1} \ell_t(\mathbf{u})$$

$$\sum_{t=0}^{T-1} \ell_t(\mathbf{w}_{t+1}) \leq \sum_{t=0}^{T-1} \ell_t(\mathbf{w}_{T+1})$$

$$\sum_{t=0}^T \ell_t(\mathbf{w}_{t+1}) \leq \sum_{t=0}^T \ell_t(\mathbf{w}_{T+1}) = \inf_{\mathbf{u} \in \mathbb{V}} \sum_{t=0}^T \ell_t(\mathbf{u})$$

(we prove this by induction on T)

(base case $T = 0$: $\mathbf{w}_1 = \operatorname{argmin}_{\mathbf{w} \in \mathbb{V}} \ell_0(\mathbf{w})$)

($T - 1 \rightarrow T$)

(choose $\mathbf{u} = \mathbf{w}_{T+1}$)

(add $\ell_T(\mathbf{w}_{T+1})$ on both sides)

FTRL stability lemma

$$\sum_{t=0}^T \ell_t(\mathbf{w}_{t+1}) \leq \inf_{\mathbf{u} \in \mathbb{V}} \sum_{t=0}^T \ell_t(\mathbf{u})$$

(we prove this by induction on T)

$$\ell_0(\mathbf{w}_1) \leq \inf_{\mathbf{u} \in \mathbb{V}} \ell_0(\mathbf{u})$$

(base case $T = 0$: $\mathbf{w}_1 = \operatorname{argmin}_{\mathbf{w} \in \mathbb{V}} \ell_0(\mathbf{w})$)

$$\sum_{t=0}^{T-1} \ell_t(\mathbf{w}_{t+1}) \leq \inf_{\mathbf{u} \in \mathbb{V}} \sum_{t=0}^{T-1} \ell_t(\mathbf{u})$$

($T - 1 \rightarrow T$)

$$\sum_{t=0}^{T-1} \ell_t(\mathbf{w}_{t+1}) \leq \sum_{t=0}^{T-1} \ell_t(\mathbf{w}_{T+1})$$

(choose $\mathbf{u} = \mathbf{w}_{T+1}$)

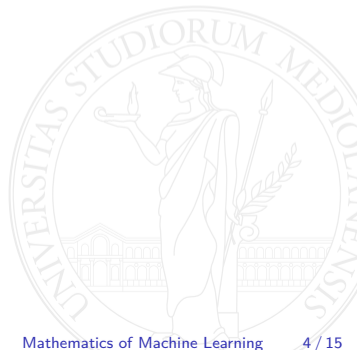
$$\sum_{t=0}^T \ell_t(\mathbf{w}_{t+1}) \leq \sum_{t=0}^T \ell_t(\mathbf{w}_{T+1}) = \inf_{\mathbf{u} \in \mathbb{V}} \sum_{t=0}^T \ell_t(\mathbf{u})$$

(add $\ell_T(\mathbf{w}_{T+1})$ on both sides)

$$R_T = \sum_{t=1}^T \left(\ell_t(\mathbf{w}_t) - \ell_t(\mathbf{w}^*) \right) \leq \ell_0(\mathbf{w}^*) - \ell_0(\mathbf{w}_1) + \sum_{t=1}^T \left(\ell_t(\mathbf{w}_t) - \ell_t(\mathbf{w}_{t+1}) \right)$$

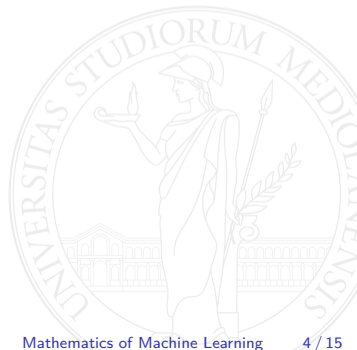
FTRL stability analysis

- ▶ Assume ψ is μ -strongly convex with respect to $\|\cdot\|$



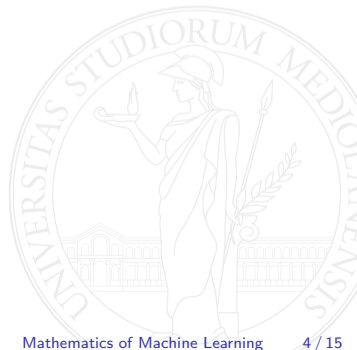
FTRL stability analysis

- ▶ Assume ψ is μ -strongly convex with respect to $\|\cdot\|$
- ▶ Pick a learning rate $\eta > 0$ and consider $\frac{\psi}{\eta}$ which is μ/η -strongly convex



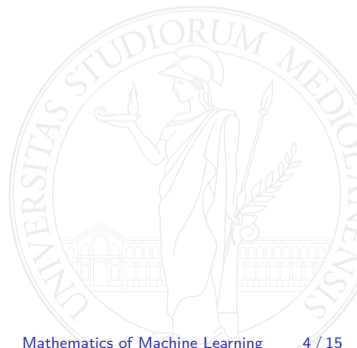
FTRL stability analysis

- ▶ Assume ψ is μ -strongly convex with respect to $\|\cdot\|$
- ▶ Pick a learning rate $\eta > 0$ and consider $\frac{\psi}{\eta}$ which is μ/η -strongly convex
- ▶ For all $t \geq 1$, we assume ℓ_t is G -Lipschitz with respect to $\|\cdot\|$



FTRL stability analysis

- ▶ Assume ψ is μ -strongly convex with respect to $\|\cdot\|$
- ▶ Pick a learning rate $\eta > 0$ and consider $\frac{\psi}{\eta}$ which is μ/η -strongly convex
- ▶ For all $t \geq 1$, we assume ℓ_t is G -Lipschitz with respect to $\|\cdot\|$
- ▶ $\mathbf{w}_{t+1} = \underset{\mathbf{w} \in \mathcal{V}}{\operatorname{argmin}} L_t(\mathbf{w})$



FTRL stability analysis

- ▶ Assume ψ is μ -strongly convex with respect to $\|\cdot\|$
- ▶ Pick a learning rate $\eta > 0$ and consider $\frac{\psi}{\eta}$ which is μ/η -strongly convex
- ▶ For all $t \geq 1$, we assume ℓ_t is G -Lipschitz with respect to $\|\cdot\|$
- ▶ $\mathbf{w}_{t+1} = \operatorname{argmin}_{\mathbf{w} \in \mathcal{V}} L_t(\mathbf{w})$

$$L_{t-1}(\mathbf{w}_{t+1}) - L_{t-1}(\mathbf{w}_t) \geq \nabla L_{t-1}(\mathbf{w}_t)^\top (\mathbf{w}_{t+1} - \mathbf{w}_t) + \frac{\mu}{2\eta} \|\mathbf{w}_t - \mathbf{w}_{t+1}\|^2 \geq \frac{\mu}{2\eta} \|\mathbf{w}_t - \mathbf{w}_{t+1}\|^2$$

FTRL stability analysis

- ▶ Assume ψ is μ -strongly convex with respect to $\|\cdot\|$
- ▶ Pick a learning rate $\eta > 0$ and consider $\frac{\psi}{\eta}$ which is μ/η -strongly convex
- ▶ For all $t \geq 1$, we assume ℓ_t is G -Lipschitz with respect to $\|\cdot\|$
- ▶ $\mathbf{w}_{t+1} = \underset{\mathbf{w} \in \mathcal{V}}{\operatorname{argmin}} L_t(\mathbf{w})$

$$L_{t-1}(\mathbf{w}_{t+1}) - L_{t-1}(\mathbf{w}_t) \geq \nabla L_{t-1}(\mathbf{w}_t)^\top (\mathbf{w}_{t+1} - \mathbf{w}_t) + \frac{\mu}{2\eta} \|\mathbf{w}_t - \mathbf{w}_{t+1}\|^2 \geq \frac{\mu}{2\eta} \|\mathbf{w}_t - \mathbf{w}_{t+1}\|^2$$

$$L_t(\mathbf{w}_t) - L_t(\mathbf{w}_{t+1}) \geq \nabla L_t(\mathbf{w}_{t+1})^\top (\mathbf{w}_t - \mathbf{w}_{t+1}) + \frac{\mu}{2\eta} \|\mathbf{w}_t - \mathbf{w}_{t+1}\|^2 \geq \frac{\mu}{2\eta} \|\mathbf{w}_t - \mathbf{w}_{t+1}\|^2$$

FTRL stability analysis

- ▶ Assume ψ is μ -strongly convex with respect to $\|\cdot\|$
- ▶ Pick a learning rate $\eta > 0$ and consider $\frac{\psi}{\eta}$ which is μ/η -strongly convex
- ▶ For all $t \geq 1$, we assume ℓ_t is G -Lipschitz with respect to $\|\cdot\|$
- ▶ $\mathbf{w}_{t+1} = \operatorname{argmin}_{\mathbf{w} \in \mathcal{V}} L_t(\mathbf{w})$

$$L_{t-1}(\mathbf{w}_{t+1}) - L_{t-1}(\mathbf{w}_t) \geq \nabla L_{t-1}(\mathbf{w}_t)^\top (\mathbf{w}_{t+1} - \mathbf{w}_t) + \frac{\mu}{2\eta} \|\mathbf{w}_t - \mathbf{w}_{t+1}\|^2 \geq \frac{\mu}{2\eta} \|\mathbf{w}_t - \mathbf{w}_{t+1}\|^2$$

$$L_t(\mathbf{w}_t) - L_t(\mathbf{w}_{t+1}) \geq \nabla L_t(\mathbf{w}_{t+1})^\top (\mathbf{w}_t - \mathbf{w}_{t+1}) + \frac{\mu}{2\eta} \|\mathbf{w}_t - \mathbf{w}_{t+1}\|^2 \geq \frac{\mu}{2\eta} \|\mathbf{w}_t - \mathbf{w}_{t+1}\|^2$$

$$\ell_t(\mathbf{w}_t) - \ell_t(\mathbf{w}_{t+1}) \geq \frac{\mu}{\eta} \|\mathbf{w}_t - \mathbf{w}_{t+1}\|^2 \quad (\text{by summing the two above inequalities})$$

FTRL stability analysis

- ▶ Assume ψ is μ -strongly convex with respect to $\|\cdot\|$
- ▶ Pick a learning rate $\eta > 0$ and consider $\frac{\psi}{\eta}$ which is μ/η -strongly convex
- ▶ For all $t \geq 1$, we assume ℓ_t is G -Lipschitz with respect to $\|\cdot\|$
- ▶ $\mathbf{w}_{t+1} = \underset{\mathbf{w} \in \mathbb{V}}{\operatorname{argmin}} L_t(\mathbf{w})$

$$L_{t-1}(\mathbf{w}_{t+1}) - L_{t-1}(\mathbf{w}_t) \geq \nabla L_{t-1}(\mathbf{w}_t)^\top (\mathbf{w}_{t+1} - \mathbf{w}_t) + \frac{\mu}{2\eta} \|\mathbf{w}_t - \mathbf{w}_{t+1}\|^2 \geq \frac{\mu}{2\eta} \|\mathbf{w}_t - \mathbf{w}_{t+1}\|^2$$

$$L_t(\mathbf{w}_t) - L_t(\mathbf{w}_{t+1}) \geq \nabla L_t(\mathbf{w}_{t+1})^\top (\mathbf{w}_t - \mathbf{w}_{t+1}) + \frac{\mu}{2\eta} \|\mathbf{w}_t - \mathbf{w}_{t+1}\|^2 \geq \frac{\mu}{2\eta} \|\mathbf{w}_t - \mathbf{w}_{t+1}\|^2$$

$$\ell_t(\mathbf{w}_t) - \ell_t(\mathbf{w}_{t+1}) \geq \frac{\mu}{\eta} \|\mathbf{w}_t - \mathbf{w}_{t+1}\|^2 \quad (\text{by summing the two above inequalities})$$

$$\ell_t(\mathbf{w}_t) - \ell_t(\mathbf{w}_{t+1}) \leq G \|\mathbf{w}_t - \mathbf{w}_{t+1}\| \quad (G\text{-Lipschitzness of } \ell_t \text{ for } t \geq 1)$$

FTRL stability analysis

- ▶ Assume ψ is μ -strongly convex with respect to $\|\cdot\|$
- ▶ Pick a learning rate $\eta > 0$ and consider $\frac{\psi}{\eta}$ which is μ/η -strongly convex
- ▶ For all $t \geq 1$, we assume ℓ_t is G -Lipschitz with respect to $\|\cdot\|$
- ▶ $\mathbf{w}_{t+1} = \underset{\mathbf{w} \in \mathbb{V}}{\operatorname{argmin}} L_t(\mathbf{w})$

$$L_{t-1}(\mathbf{w}_{t+1}) - L_{t-1}(\mathbf{w}_t) \geq \nabla L_{t-1}(\mathbf{w}_t)^\top (\mathbf{w}_{t+1} - \mathbf{w}_t) + \frac{\mu}{2\eta} \|\mathbf{w}_t - \mathbf{w}_{t+1}\|^2 \geq \frac{\mu}{2\eta} \|\mathbf{w}_t - \mathbf{w}_{t+1}\|^2$$

$$L_t(\mathbf{w}_t) - L_t(\mathbf{w}_{t+1}) \geq \nabla L_t(\mathbf{w}_{t+1})^\top (\mathbf{w}_t - \mathbf{w}_{t+1}) + \frac{\mu}{2\eta} \|\mathbf{w}_t - \mathbf{w}_{t+1}\|^2 \geq \frac{\mu}{2\eta} \|\mathbf{w}_t - \mathbf{w}_{t+1}\|^2$$

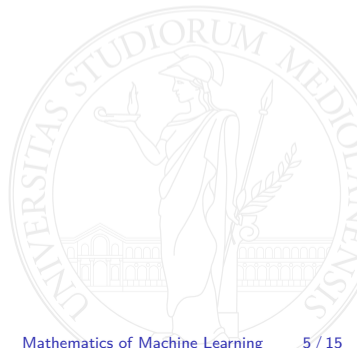
$$\ell_t(\mathbf{w}_t) - \ell_t(\mathbf{w}_{t+1}) \geq \frac{\mu}{\eta} \|\mathbf{w}_t - \mathbf{w}_{t+1}\|^2 \quad (\text{by summing the two above inequalities})$$

$$\ell_t(\mathbf{w}_t) - \ell_t(\mathbf{w}_{t+1}) \leq G \|\mathbf{w}_t - \mathbf{w}_{t+1}\| \quad (G\text{-Lipschitzness of } \ell_t \text{ for } t \geq 1)$$

$$\|\mathbf{w}_t - \mathbf{w}_{t+1}\| \leq \frac{G\eta}{\mu} \quad \text{implying} \quad \ell_t(\mathbf{w}_t) - \ell_t(\mathbf{w}_{t+1}) \leq \frac{G^2\eta}{\mu}$$

FTRL regret bound

$$R_T \leq \frac{\psi(\mathbf{w}^*) - \psi(\mathbf{w}_1)}{\eta} + \sum_{t=1}^T (\ell_t(\mathbf{w}_t) - \ell_t(\mathbf{w}_{t+1})) \leq \frac{\psi(\mathbf{w}^*) - \psi(\mathbf{w}_1)}{\eta} + \eta \frac{G^2}{\mu} T$$

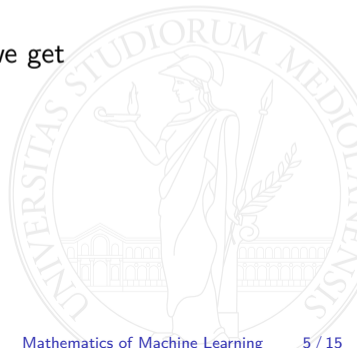


FTRL regret bound

$$R_T \leq \frac{\psi(\mathbf{w}^*) - \psi(\mathbf{w}_1)}{\eta} + \sum_{t=1}^T \left(\ell_t(\mathbf{w}_t) - \ell_t(\mathbf{w}_{t+1}) \right) \leq \frac{\psi(\mathbf{w}^*) - \psi(\mathbf{w}_1)}{\eta} + \eta \frac{G^2}{\mu} T$$

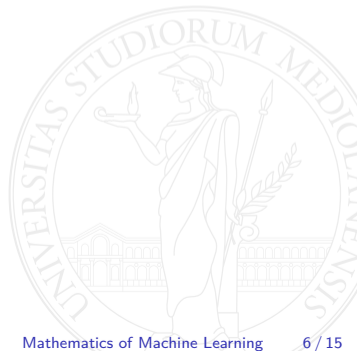
Assuming $\max_{\mathbf{u} \in \mathcal{V}} \psi(\mathbf{u}) - \min_{\mathbf{w} \in \mathcal{V}} \psi(\mathbf{w}) = D^2$ and choosing $\eta = \frac{D}{G} \sqrt{\frac{\mu}{T}}$ we get

$$R_T \leq 2GD \sqrt{\frac{T}{\mu}} = \mathcal{O}(GD\sqrt{T})$$



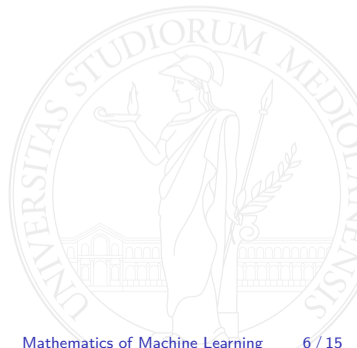
Lipschitz constants and dual norm of gradients

- ▶ Dual norm of $\|\cdot\|$ is $\|\mathbf{w}\|_* = \sup_{\mathbf{u} \in \mathbb{R}^d} \frac{\mathbf{w}^\top \mathbf{u}}{\|\mathbf{u}\|}$



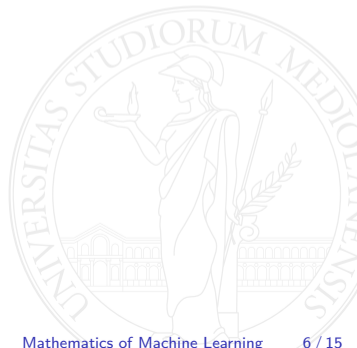
Lipschitz constants and dual norm of gradients

- ▶ Dual norm of $\|\cdot\|$ is $\|\mathbf{w}\|_* = \sup_{\mathbf{u} \in \mathbb{R}^d} \frac{\mathbf{w}^\top \mathbf{u}}{\|\mathbf{u}\|}$
- ▶ Example (p -norms): the dual norm of $\|\cdot\|_p$ is $\|\cdot\|_q$ where $\frac{1}{p} + \frac{1}{q} = 1$ and $p, q > 1$



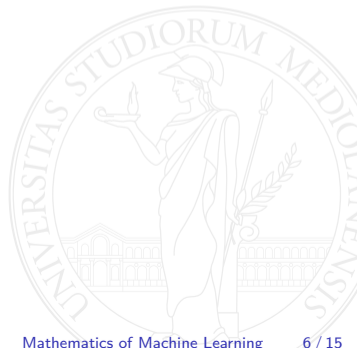
Lipschitz constants and dual norm of gradients

- ▶ Dual norm of $\|\cdot\|$ is $\|\mathbf{w}\|_* = \sup_{\mathbf{u} \in \mathbb{R}^d} \frac{\mathbf{w}^\top \mathbf{u}}{\|\mathbf{u}\|}$
- ▶ Example (p -norms): the dual norm of $\|\cdot\|_p$ is $\|\cdot\|_q$ where $\frac{1}{p} + \frac{1}{q} = 1$ and $p, q > 1$
- ▶ Relationship to convex duality: $\psi = \frac{1}{2} \|\cdot\|_p^2$ and $\psi^* = \frac{1}{2} \|\cdot\|_q^2$



Lipschitz constants and dual norm of gradients

- ▶ Dual norm of $\|\cdot\|$ is $\|\mathbf{w}\|_* = \sup_{\mathbf{u} \in \mathbb{R}^d} \frac{\mathbf{w}^\top \mathbf{u}}{\|\mathbf{u}\|}$
- ▶ Example (p -norms): the dual norm of $\|\cdot\|_p$ is $\|\cdot\|_q$ where $\frac{1}{p} + \frac{1}{q} = 1$ and $p, q > 1$
- ▶ Relationship to convex duality: $\psi = \frac{1}{2} \|\cdot\|_p^2$ and $\psi^* = \frac{1}{2} \|\cdot\|_q^2$
- ▶ Hölder inequality: $\mathbf{u}^\top \mathbf{w} \leq \|\mathbf{u}\| \|\mathbf{w}\|_*$
- ▶ Fenchel-Young inequality: $\mathbf{u}^\top \mathbf{w} \leq \psi(\mathbf{u}) + \psi^*(\mathbf{w})$



Lipschitz constants and dual norm of gradients

- ▶ Dual norm of $\|\cdot\|$ is $\|\mathbf{w}\|_* = \sup_{\mathbf{u} \in \mathbb{R}^d} \frac{\mathbf{w}^\top \mathbf{u}}{\|\mathbf{u}\|}$
- ▶ Example (p -norms): the dual norm of $\|\cdot\|_p$ is $\|\cdot\|_q$ where $\frac{1}{p} + \frac{1}{q} = 1$ and $p, q > 1$
- ▶ Relationship to convex duality: $\psi = \frac{1}{2} \|\cdot\|_p^2$ and $\psi^* = \frac{1}{2} \|\cdot\|_q^2$
- ▶ Hölder inequality: $\mathbf{u}^\top \mathbf{w} \leq \|\mathbf{u}\| \|\mathbf{w}\|_*$
- ▶ Fenchel-Young inequality: $\mathbf{u}^\top \mathbf{w} \leq \psi(\mathbf{u}) + \psi^*(\mathbf{w})$

Theorem

Let $\ell : \mathbb{V} \rightarrow \mathbb{R}$ be a differentiable convex function. Then ℓ is G -Lipschitz over \mathbb{V} with respect to a norm $\|\cdot\|$ iff for all $\mathbf{w} \in \mathbb{V}$ we have that $\|\nabla \ell(\mathbf{w})\|_* \leq G$

Lipschitz constants and dual norm of gradients

- ▶ Dual norm of $\|\cdot\|$ is $\|\mathbf{w}\|_* = \sup_{\mathbf{u} \in \mathbb{R}^d} \frac{\mathbf{w}^\top \mathbf{u}}{\|\mathbf{u}\|}$
- ▶ Example (p -norms): the dual norm of $\|\cdot\|_p$ is $\|\cdot\|_q$ where $\frac{1}{p} + \frac{1}{q} = 1$ and $p, q > 1$
- ▶ Relationship to convex duality: $\psi = \frac{1}{2} \|\cdot\|_p^2$ and $\psi^* = \frac{1}{2} \|\cdot\|_q^2$
- ▶ Hölder inequality: $\mathbf{u}^\top \mathbf{w} \leq \|\mathbf{u}\| \|\mathbf{w}\|_*$
- ▶ Fenchel-Young inequality: $\mathbf{u}^\top \mathbf{w} \leq \psi(\mathbf{u}) + \psi^*(\mathbf{w})$

Theorem

Let $\ell : \mathbb{V} \rightarrow \mathbb{R}$ be a differentiable convex function. Then ℓ is G -Lipschitz over \mathbb{V} with respect to a norm $\|\cdot\|$ iff for all $\mathbf{w} \in \mathbb{V}$ we have that $\|\nabla \ell(\mathbf{w})\|_* \leq G$

- ▶ Assume $\max_i |(\nabla \ell)_i| = \Theta(1)$

Lipschitz constants and dual norm of gradients

- ▶ Dual norm of $\|\cdot\|$ is $\|\mathbf{w}\|_* = \sup_{\mathbf{u} \in \mathbb{R}^d} \frac{\mathbf{w}^\top \mathbf{u}}{\|\mathbf{u}\|}$
- ▶ Example (p -norms): the dual norm of $\|\cdot\|_p$ is $\|\cdot\|_q$ where $\frac{1}{p} + \frac{1}{q} = 1$ and $p, q > 1$
- ▶ Relationship to convex duality: $\psi = \frac{1}{2} \|\cdot\|_p^2$ and $\psi^* = \frac{1}{2} \|\cdot\|_q^2$
- ▶ Hölder inequality: $\mathbf{u}^\top \mathbf{w} \leq \|\mathbf{u}\| \|\mathbf{w}\|_*$
- ▶ Fenchel-Young inequality: $\mathbf{u}^\top \mathbf{w} \leq \psi(\mathbf{u}) + \psi^*(\mathbf{w})$

Theorem

Let $\ell : \mathbb{V} \rightarrow \mathbb{R}$ be a differentiable convex function. Then ℓ is G -Lipschitz over \mathbb{V} with respect to a norm $\|\cdot\|$ iff for all $\mathbf{w} \in \mathbb{V}$ we have that $\|\nabla \ell(\mathbf{w})\|_* \leq G$

- ▶ Assume $\max_i |(\nabla \ell)_i| = \Theta(1)$
- ▶ If ℓ is G -Lipschitz with respect to $\|\cdot\|_2$, then $G = \mathcal{O}(\sqrt{d})$

Lipschitz constants and dual norm of gradients

- ▶ Dual norm of $\|\cdot\|$ is $\|\mathbf{w}\|_* = \sup_{\mathbf{u} \in \mathbb{R}^d} \frac{\mathbf{w}^\top \mathbf{u}}{\|\mathbf{u}\|}$
- ▶ Example (p -norms): the dual norm of $\|\cdot\|_p$ is $\|\cdot\|_q$ where $\frac{1}{p} + \frac{1}{q} = 1$ and $p, q > 1$
- ▶ Relationship to convex duality: $\psi = \frac{1}{2} \|\cdot\|_p^2$ and $\psi^* = \frac{1}{2} \|\cdot\|_q^2$
- ▶ Hölder inequality: $\mathbf{u}^\top \mathbf{w} \leq \|\mathbf{u}\| \|\mathbf{w}\|_*$
- ▶ Fenchel-Young inequality: $\mathbf{u}^\top \mathbf{w} \leq \psi(\mathbf{u}) + \psi^*(\mathbf{w})$

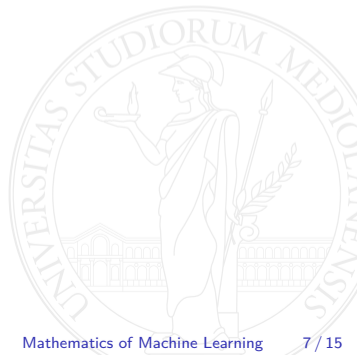
Theorem

Let $\ell : \mathbb{V} \rightarrow \mathbb{R}$ be a differentiable convex function. Then ℓ is G -Lipschitz over \mathbb{V} with respect to a norm $\|\cdot\|$ iff for all $\mathbf{w} \in \mathbb{V}$ we have that $\|\nabla \ell(\mathbf{w})\|_* \leq G$

- ▶ Assume $\max_i |(\nabla \ell)_i| = \Theta(1)$
- ▶ If ℓ is G -Lipschitz with respect to $\|\cdot\|_2$, then $G = \mathcal{O}(\sqrt{d})$
- ▶ If ℓ is G -Lipschitz with respect to $\|\cdot\|_1$, then $G = \mathcal{O}(1)$

Matching the regularizer to the geometry of the model space

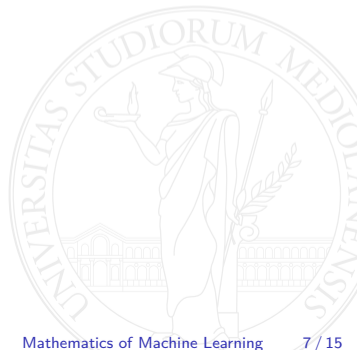
Projected Lazy OGD



Matching the regularizer to the geometry of the model space

Projected Lazy OGD

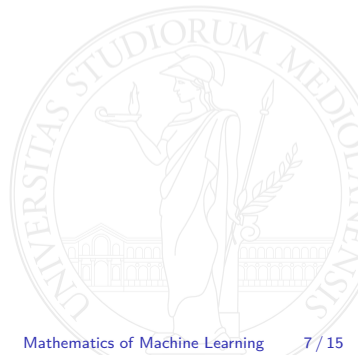
- ▶ Take \mathbb{V} to be the closed Euclidean ball of radius D



Matching the regularizer to the geometry of the model space

Projected Lazy OGD

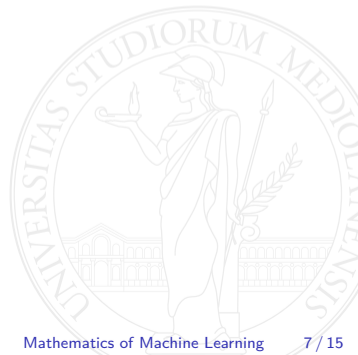
- ▶ Take \mathbb{V} to be the closed Euclidean ball of radius D
- ▶ $\psi = \frac{1}{2} \|\cdot\|_2^2$ is 1-strongly convex with respect to $\|\cdot\|_2$



Matching the regularizer to the geometry of the model space

Projected Lazy OGD

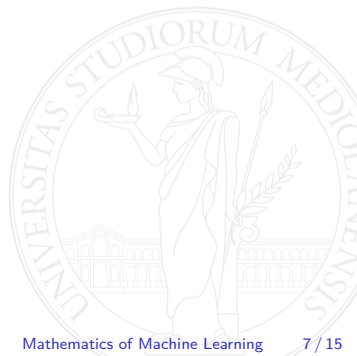
- ▶ Take \mathbb{V} to be the closed Euclidean ball of radius D
- ▶ $\psi = \frac{1}{2} \|\cdot\|_2^2$ is 1-strongly convex with respect to $\|\cdot\|_2$
- ▶ $\max_{\mathbf{u} \in \mathbb{V}} \psi(\mathbf{u}) - \min_{\mathbf{w} \in \mathbb{V}} \psi(\mathbf{w}) = \frac{D^2}{2}$



Matching the regularizer to the geometry of the model space

Projected Lazy OGD

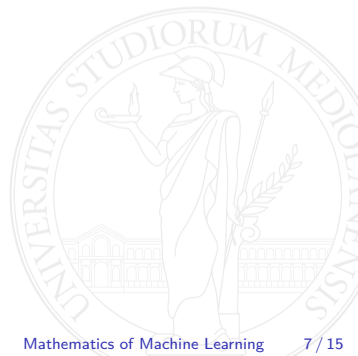
- ▶ Take \mathbb{V} to be the closed Euclidean ball of radius D
- ▶ $\psi = \frac{1}{2} \|\cdot\|_2^2$ is 1-strongly convex with respect to $\|\cdot\|_2$
- ▶ $\max_{\mathbf{u} \in \mathbb{V}} \psi(\mathbf{u}) - \min_{\mathbf{w} \in \mathbb{V}} \psi(\mathbf{w}) = \frac{D^2}{2}$
- ▶ Lipschitz constant: $G = \mathcal{O}(\sqrt{d})$



Matching the regularizer to the geometry of the model space

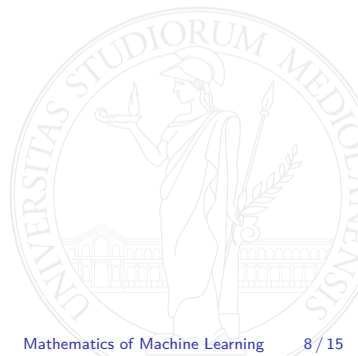
Projected Lazy OGD

- ▶ Take \mathbb{V} to be the closed Euclidean ball of radius D
- ▶ $\psi = \frac{1}{2} \|\cdot\|_2^2$ is 1-strongly convex with respect to $\|\cdot\|_2$
- ▶ $\max_{\mathbf{u} \in \mathbb{V}} \psi(\mathbf{u}) - \min_{\mathbf{w} \in \mathbb{V}} \psi(\mathbf{w}) = \frac{D^2}{2}$
- ▶ Lipschitz constant: $G = \mathcal{O}(\sqrt{d})$
- ▶ $R_T \leq 2GD\sqrt{\frac{T}{\mu}} = \mathcal{O}(D\sqrt{dT})$



Matching the regularizer to the geometry of the model space

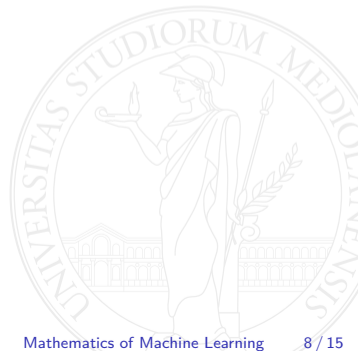
EG



Matching the regularizer to the geometry of the model space

EG

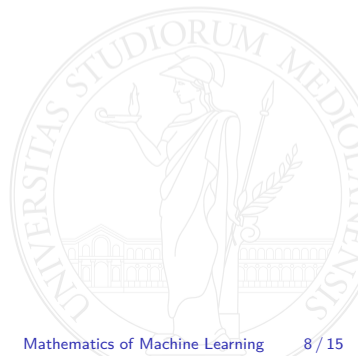
- ▶ \mathbb{V} is probability simplex Δ_d



Matching the regularizer to the geometry of the model space

EG

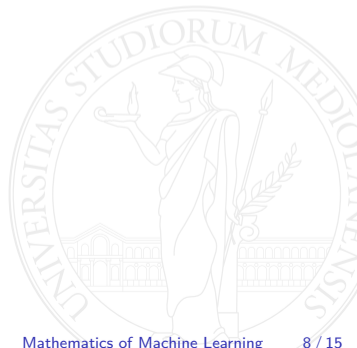
- ▶ \mathbb{V} is probability simplex Δ_d
- ▶ $\psi(\mathbf{p}) = \sum_i p_i \ln p_i$ is 1-strongly convex with respect to $\|\cdot\|_1$



Matching the regularizer to the geometry of the model space

EG

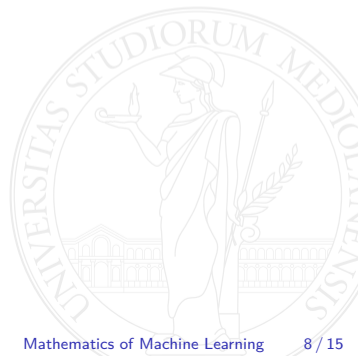
- ▶ \mathbb{V} is probability simplex Δ_d
- ▶ $\psi(\mathbf{p}) = \sum_i p_i \ln p_i$ is 1-strongly convex with respect to $\|\cdot\|_1$
- ▶ $D = \max_{\mathbf{u} \in \mathbb{V}} \psi(\mathbf{u}) - \min_{\mathbf{w} \in \mathbb{V}} \psi(\mathbf{w}) = \ln d$



Matching the regularizer to the geometry of the model space

EG

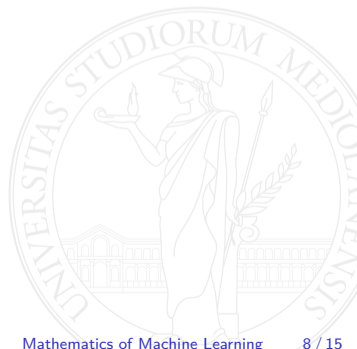
- ▶ \mathbb{V} is probability simplex Δ_d
- ▶ $\psi(\mathbf{p}) = \sum_i p_i \ln p_i$ is 1-strongly convex with respect to $\|\cdot\|_1$
- ▶ $D = \max_{\mathbf{u} \in \mathbb{V}} \psi(\mathbf{u}) - \min_{\mathbf{w} \in \mathbb{V}} \psi(\mathbf{w}) = \ln d$
- ▶ Lipschitz constant: $G = \mathcal{O}(1)$



Matching the regularizer to the geometry of the model space

EG

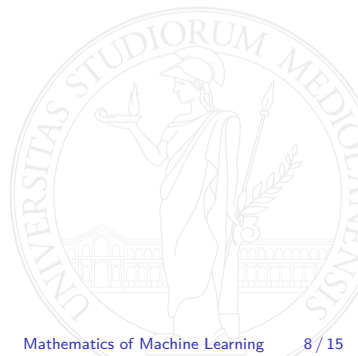
- ▶ \mathbb{V} is probability simplex Δ_d
- ▶ $\psi(\mathbf{p}) = \sum_i p_i \ln p_i$ is 1-strongly convex with respect to $\|\cdot\|_1$
- ▶ $D = \max_{\mathbf{u} \in \mathbb{V}} \psi(\mathbf{u}) - \min_{\mathbf{w} \in \mathbb{V}} \psi(\mathbf{w}) = \ln d$
- ▶ Lipschitz constant: $G = \mathcal{O}(1)$
- ▶ $R_T \leq 2GD\sqrt{\frac{T}{\mu}} = \mathcal{O}(\sqrt{T \ln d})$



Matching the regularizer to the geometry of the model space

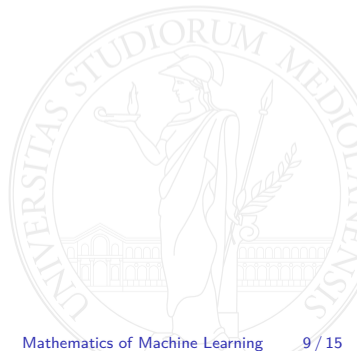
EG

- ▶ \mathbb{V} is probability simplex Δ_d
- ▶ $\psi(\mathbf{p}) = \sum_i p_i \ln p_i$ is 1-strongly convex with respect to $\|\cdot\|_1$
- ▶ $D = \max_{\mathbf{u} \in \mathbb{V}} \psi(\mathbf{u}) - \min_{\mathbf{w} \in \mathbb{V}} \psi(\mathbf{w}) = \ln d$
- ▶ Lipschitz constant: $G = \mathcal{O}(1)$
- ▶ $R_T \leq 2GD\sqrt{\frac{T}{\mu}} = \mathcal{O}(\sqrt{T \ln d})$
- ▶ For $\mathbb{V} = \Delta_d$, projected lazy OGD only achieves $R_T = \mathcal{O}(\sqrt{dT})$
- ▶ The geometry of \mathbb{V} matters



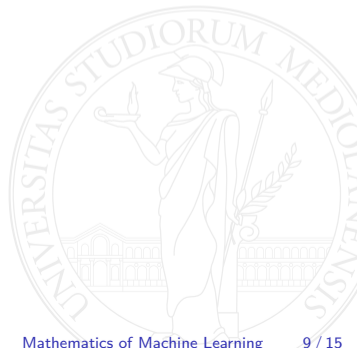
Lower bounds

- ▶ V is a bounded set of Euclidean diameter D



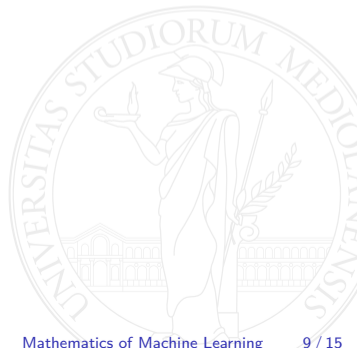
Lower bounds

- ▶ \mathbb{V} is a bounded set of Euclidean diameter D
- ▶ Take $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{V}$ such that $\|\mathbf{v}_1 - \mathbf{v}_2\|_2 = D$ and set $\mathbf{z}_0 = (\mathbf{v}_1 - \mathbf{v}_2) / \|\mathbf{v}_1 - \mathbf{v}_2\|_2$



Lower bounds

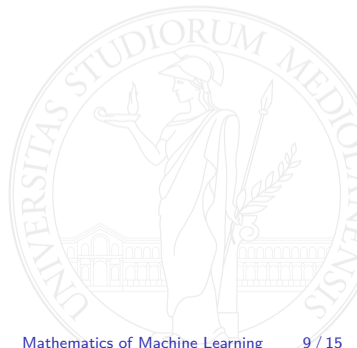
- ▶ \mathbb{V} is a bounded set of Euclidean diameter D
- ▶ Take $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{V}$ such that $\|\mathbf{v}_1 - \mathbf{v}_2\|_2 = D$ and set $\mathbf{z}_0 = (\mathbf{v}_1 - \mathbf{v}_2) / \|\mathbf{v}_1 - \mathbf{v}_2\|_2$
- ▶ G -Lipschitz **linear losses**: $l_t(\mathbf{w}) = \varepsilon_t G \mathbf{w}^\top \mathbf{z}_0$ where $\varepsilon_t \in \{-1, 1\}$ are uniform i.i.d.



Lower bounds

- ▶ \mathbb{V} is a bounded set of Euclidean diameter D
- ▶ Take $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{V}$ such that $\|\mathbf{v}_1 - \mathbf{v}_2\|_2 = D$ and set $\mathbf{z}_0 = (\mathbf{v}_1 - \mathbf{v}_2) / \|\mathbf{v}_1 - \mathbf{v}_2\|_2$
- ▶ G -Lipschitz **linear losses**: $l_t(\mathbf{w}) = \varepsilon_t G \mathbf{w}^\top \mathbf{z}_0$ where $\varepsilon_t \in \{-1, 1\}$ are uniform i.i.d.

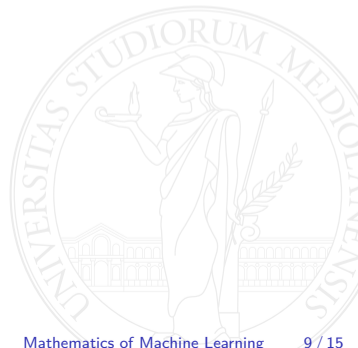
$$\mathbb{E} \left[\max_{\mathbf{u} \in \{\mathbf{v}_1, \mathbf{v}_2\}} R_T(\mathbf{u}) \right]$$



Lower bounds

- ▶ \mathbb{V} is a bounded set of Euclidean diameter D
- ▶ Take $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{V}$ such that $\|\mathbf{v}_1 - \mathbf{v}_2\|_2 = D$ and set $\mathbf{z}_0 = (\mathbf{v}_1 - \mathbf{v}_2) / \|\mathbf{v}_1 - \mathbf{v}_2\|_2$
- ▶ G -Lipschitz **linear losses**: $\ell_t(\mathbf{w}) = \varepsilon_t G \mathbf{w}^\top \mathbf{z}_0$ where $\varepsilon_t \in \{-1, 1\}$ are uniform i.i.d.

$$\mathbb{E} \left[\max_{\mathbf{u} \in \{\mathbf{v}_1, \mathbf{v}_2\}} R_T(\mathbf{u}) \right] = \mathbb{E} \left[\max_{\mathbf{u} \in \{\mathbf{v}_1, \mathbf{v}_2\}} \sum_{t=1}^T \ell_t(\mathbf{u}) \right] \quad (\text{since } \mathbb{E}[\ell_t(\mathbf{w})] = 0)$$



Lower bounds

- ▶ \mathbb{V} is a bounded set of Euclidean diameter D
- ▶ Take $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{V}$ such that $\|\mathbf{v}_1 - \mathbf{v}_2\|_2 = D$ and set $\mathbf{z}_0 = (\mathbf{v}_1 - \mathbf{v}_2) / \|\mathbf{v}_1 - \mathbf{v}_2\|_2$
- ▶ G -Lipschitz **linear losses**: $\ell_t(\mathbf{w}) = \varepsilon_t G \mathbf{w}^\top \mathbf{z}_0$ where $\varepsilon_t \in \{-1, 1\}$ are uniform i.i.d.

$$\begin{aligned} \mathbb{E} \left[\max_{\mathbf{u} \in \{\mathbf{v}_1, \mathbf{v}_2\}} R_T(\mathbf{u}) \right] &= \mathbb{E} \left[\max_{\mathbf{u} \in \{\mathbf{v}_1, \mathbf{v}_2\}} \sum_{t=1}^T \ell_t(\mathbf{u}) \right] && \text{(since } \mathbb{E}[\ell_t(\mathbf{w})] = 0 \text{)} \\ &= \frac{G}{2} \mathbb{E} \left[\left| \sum_{t=1}^T \varepsilon_t \mathbf{z}_0^\top (\mathbf{v}_1 - \mathbf{v}_2) \right| \right] && \text{(using } \max\{a, b\} = \frac{1}{2}(a + b + |a - b|) \text{)} \end{aligned}$$

Lower bounds

- ▶ \mathbb{V} is a bounded set of Euclidean diameter D
- ▶ Take $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{V}$ such that $\|\mathbf{v}_1 - \mathbf{v}_2\|_2 = D$ and set $\mathbf{z}_0 = (\mathbf{v}_1 - \mathbf{v}_2) / \|\mathbf{v}_1 - \mathbf{v}_2\|_2$
- ▶ G -Lipschitz **linear losses**: $\ell_t(\mathbf{w}) = \varepsilon_t G \mathbf{w}^\top \mathbf{z}_0$ where $\varepsilon_t \in \{-1, 1\}$ are uniform i.i.d.

$$\begin{aligned}\mathbb{E} \left[\max_{\mathbf{u} \in \{\mathbf{v}_1, \mathbf{v}_2\}} R_T(\mathbf{u}) \right] &= \mathbb{E} \left[\max_{\mathbf{u} \in \{\mathbf{v}_1, \mathbf{v}_2\}} \sum_{t=1}^T \ell_t(\mathbf{u}) \right] && \text{(since } \mathbb{E}[\ell_t(\mathbf{w})] = 0\text{)} \\ &= \frac{G}{2} \mathbb{E} \left[\left| \sum_{t=1}^T \varepsilon_t \mathbf{z}_0^\top (\mathbf{v}_1 - \mathbf{v}_2) \right| \right] && \text{(using } \max\{a, b\} = \frac{1}{2}(a + b + |a - b|)\text{)} \\ &= \frac{GD}{2} \mathbb{E} \left[\left| \sum_{t=1}^T \varepsilon_t \right| \right] && \text{(because } \mathbf{z}_0^\top (\mathbf{v}_1 - \mathbf{v}_2) = D\text{)}\end{aligned}$$

Lower bounds

- ▶ \mathbb{V} is a bounded set of Euclidean diameter D
- ▶ Take $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{V}$ such that $\|\mathbf{v}_1 - \mathbf{v}_2\|_2 = D$ and set $\mathbf{z}_0 = (\mathbf{v}_1 - \mathbf{v}_2) / \|\mathbf{v}_1 - \mathbf{v}_2\|_2$
- ▶ G -Lipschitz **linear losses**: $\ell_t(\mathbf{w}) = \varepsilon_t G \mathbf{w}^\top \mathbf{z}_0$ where $\varepsilon_t \in \{-1, 1\}$ are uniform i.i.d.

$$\begin{aligned} \mathbb{E} \left[\max_{\mathbf{u} \in \{\mathbf{v}_1, \mathbf{v}_2\}} R_T(\mathbf{u}) \right] &= \mathbb{E} \left[\max_{\mathbf{u} \in \{\mathbf{v}_1, \mathbf{v}_2\}} \sum_{t=1}^T \ell_t(\mathbf{u}) \right] && \text{(since } \mathbb{E}[\ell_t(\mathbf{w})] = 0\text{)} \\ &= \frac{G}{2} \mathbb{E} \left[\left| \sum_{t=1}^T \varepsilon_t \mathbf{z}_0^\top (\mathbf{v}_1 - \mathbf{v}_2) \right| \right] && \text{(using } \max\{a, b\} = \frac{1}{2}(a + b + |a - b|)\text{)} \\ &= \frac{GD}{2} \mathbb{E} \left[\left| \sum_{t=1}^T \varepsilon_t \right| \right] && \text{(because } \mathbf{z}_0^\top (\mathbf{v}_1 - \mathbf{v}_2) = D\text{)} \\ &\geq GD \sqrt{\frac{T}{8}} && \text{(Khintchine inequality)} \end{aligned}$$

Lower bound for the simplex

- ▶ Stochastic linear losses $\ell_t = (\ell_t(1), \dots, \ell_t(d))$



Lower bound for the simplex

- ▶ Stochastic linear losses $\ell_t = (\ell_t(1), \dots, \ell_t(d))$
- ▶ $\ell_t(i) \in \{0, 1\}$ independent random coin flip for all $t \geq 1$ and $i = 1, \dots, d$



Lower bound for the simplex

- ▶ Stochastic linear losses $\ell_t = (\ell_t(1), \dots, \ell_t(d))$
- ▶ $\ell_t(i) \in \{0, 1\}$ independent random coin flip for all $t \geq 1$ and $i = 1, \dots, d$
- ▶ For any online algorithm
$$\mathbb{E} \left[\sum_{t=1}^T \mathbf{p}_t^\top \ell_t \right] = \frac{T}{2}$$



Lower bound for the simplex

- ▶ Stochastic linear losses $\ell_t = (\ell_t(1), \dots, \ell_t(d))$
- ▶ $\ell_t(i) \in \{0, 1\}$ independent random coin flip for all $t \geq 1$ and $i = 1, \dots, d$
- ▶ For any online algorithm $\mathbb{E} \left[\sum_{t=1}^T \mathbf{p}_t^\top \ell_t \right] = \frac{T}{2}$
- ▶ Then the expected regret is

$$\begin{aligned} \frac{T}{2} - \mathbb{E} \left[\min_{\mathbf{p} \in \Delta_d} \sum_{t=1}^T \mathbf{q}^\top \ell_t \right] &= \frac{T}{2} - \mathbb{E} \left[\min_{i=1, \dots, d} \sum_{t=1}^T \ell_t(i) \right] \\ &= \mathbb{E} \left[\max_{i=1, \dots, d} \sum_{t=1}^T \left(\frac{1}{2} - \ell_t(i) \right) \right] \\ &= \frac{1}{2} \mathbb{E} \left[\max_{i=1, \dots, d} \sum_{t=1}^T \varepsilon_t(i) \right] \end{aligned}$$

where $\varepsilon_t(i) \in \{-1, 1\}$ are uniform i.i.d.

Lower bound for the simplex: finishing up

$$\mathbb{E} \left[\max_{i=1, \dots, d} \sum_{t=1}^T \varepsilon_t(i) \right] = (1 - o(1)) \sqrt{2T \ln d}$$

Therefore, $R_T = \Omega(\sqrt{T \ln d})$



FTRL with time-varying regularizer

- ▶ Fix a sequence ψ_1, ψ_2, \dots of regularizers



FTRL with time-varying regularizer

- ▶ Fix a sequence ψ_1, ψ_2, \dots of regularizers
- ▶ $L_t^\psi = \psi_{t+1} + L_t = \psi_{t+1} + \ell_1 + \dots + \ell_t$



FTRL with time-varying regularizer

► Fix a sequence ψ_1, ψ_2, \dots of regularizers

► $L_t^\psi = \psi_{t+1} + L_t = \psi_{t+1} + \ell_1 + \dots + \ell_t$

► FTRL prediction $\mathbf{w}_{t+1} = \operatorname{argmin}_{\mathbf{w} \in \mathbb{V}} L_t^\psi(\mathbf{w})$ and best model $\mathbf{w}^* = \operatorname{argmin}_{\mathbf{w} \in \mathbb{V}} \sum_{t=1}^T \ell_t(\mathbf{w})$



FTRL with time-varying regularizer

► Fix a sequence ψ_1, ψ_2, \dots of regularizers

► $L_t^\psi = \psi_{t+1} + L_t = \psi_{t+1} + \ell_1 + \dots + \ell_t$

► FTRL prediction $\mathbf{w}_{t+1} = \operatorname{argmin}_{\mathbf{w} \in \mathbb{V}} L_t^\psi(\mathbf{w})$ and best model $\mathbf{w}^* = \operatorname{argmin}_{\mathbf{w} \in \mathbb{V}} \sum_{t=1}^T \ell_t(\mathbf{w})$

$$-L_T(\mathbf{w}^*) = \psi_{T+1}(\mathbf{w}^*) - L_T^\psi(\mathbf{w}^*)$$



FTRL with time-varying regularizer

- ▶ Fix a sequence ψ_1, ψ_2, \dots of regularizers

- ▶ $L_t^\psi = \psi_{t+1} + L_t = \psi_{t+1} + \ell_1 + \dots + \ell_t$

- ▶ FTRL prediction $\mathbf{w}_{t+1} = \operatorname{argmin}_{\mathbf{w} \in \mathbb{V}} L_t^\psi(\mathbf{w})$ and best model $\mathbf{w}^* = \operatorname{argmin}_{\mathbf{w} \in \mathbb{V}} \sum_{t=1}^T \ell_t(\mathbf{w})$

$$-L_T(\mathbf{w}^*) = \psi_{T+1}(\mathbf{w}^*) - L_T^\psi(\mathbf{w}^*)$$

$$-L_T(\mathbf{w}^*) = \psi_{T+1}(\mathbf{w}^*) - \underbrace{L_0^\psi(\mathbf{w}_1)}_{=\psi_1(\mathbf{w}_1)} + \underbrace{L_0^\psi(\mathbf{w}_1) - L_T^\psi(\mathbf{w}_{T+1})}_{\text{write as telescoping}} + \underbrace{L_T^\psi(\mathbf{w}_{T+1}) - L_T^\psi(\mathbf{w}^*)}_{\leq 0}$$

FTRL with time-varying regularizer

- ▶ Fix a sequence ψ_1, ψ_2, \dots of regularizers

- ▶ $L_t^\psi = \psi_{t+1} + L_t = \psi_{t+1} + \ell_1 + \dots + \ell_t$

- ▶ FTRL prediction $\mathbf{w}_{t+1} = \operatorname{argmin}_{\mathbf{w} \in \mathbb{V}} L_t^\psi(\mathbf{w})$ and best model $\mathbf{w}^* = \operatorname{argmin}_{\mathbf{w} \in \mathbb{V}} \sum_{t=1}^T \ell_t(\mathbf{w})$

$$-L_T(\mathbf{w}^*) = \psi_{T+1}(\mathbf{w}^*) - L_T^\psi(\mathbf{w}^*)$$

$$-L_T(\mathbf{w}^*) = \psi_{T+1}(\mathbf{w}^*) - \underbrace{L_0^\psi(\mathbf{w}_1)}_{=\psi_1(\mathbf{w}_1)} + \underbrace{L_0^\psi(\mathbf{w}_1) - L_T^\psi(\mathbf{w}_{T+1})}_{\text{write as telescoping}} + \underbrace{L_T^\psi(\mathbf{w}_{T+1}) - L_T^\psi(\mathbf{w}^*)}_{\leq 0}$$

$$-L_T(\mathbf{w}^*) \leq \psi_{T+1}(\mathbf{w}^*) - \psi_1(\mathbf{w}_1) + \sum_{t=1}^T \left(L_{t-1}^\psi(\mathbf{w}_t) - L_t^\psi(\mathbf{w}_{t+1}) \right)$$

FTRL with time-varying regularizer

- ▶ Fix a sequence ψ_1, ψ_2, \dots of regularizers

- ▶ $L_t^\psi = \psi_{t+1} + L_t = \psi_{t+1} + \ell_1 + \dots + \ell_t$

- ▶ FTRL prediction $\mathbf{w}_{t+1} = \operatorname{argmin}_{\mathbf{w} \in \mathbb{V}} L_t^\psi(\mathbf{w})$ and best model $\mathbf{w}^* = \operatorname{argmin}_{\mathbf{w} \in \mathbb{V}} \sum_{t=1}^T \ell_t(\mathbf{w})$

$$-L_T(\mathbf{w}^*) = \psi_{T+1}(\mathbf{w}^*) - L_T^\psi(\mathbf{w}^*)$$

$$-L_T(\mathbf{w}^*) = \psi_{T+1}(\mathbf{w}^*) - \underbrace{L_0^\psi(\mathbf{w}_1)}_{=\psi_1(\mathbf{w}_1)} + \underbrace{L_0^\psi(\mathbf{w}_1) - L_T^\psi(\mathbf{w}_{T+1})}_{\text{write as telescoping}} + \underbrace{L_T^\psi(\mathbf{w}_{T+1}) - L_T^\psi(\mathbf{w}^*)}_{\leq 0}$$

$$-L_T(\mathbf{w}^*) \leq \psi_{T+1}(\mathbf{w}^*) - \psi_1(\mathbf{w}_1) + \sum_{t=1}^T \left(L_{t-1}^\psi(\mathbf{w}_t) - L_t^\psi(\mathbf{w}_{t+1}) \right)$$

$$R_T \leq \psi_{T+1}(\mathbf{w}^*) - \psi_1(\mathbf{w}_1) + \sum_{t=1}^T \left(L_{t-1}^\psi(\mathbf{w}_t) - L_t^\psi(\mathbf{w}_{t+1}) + \ell_t(\mathbf{w}_t) \right)$$

Analysis of FTRL with time-varying regularizer

$$R_T \leq \psi_{T+1}(\mathbf{w}^*) - \psi_1(\mathbf{w}_1) + \sum_{t=1}^T \left(L_{t-1}^\psi(\mathbf{w}_t) - L_t^\psi(\mathbf{w}_{t+1}) + \ell_t(\mathbf{w}_t) \right)$$



Analysis of FTRL with time-varying regularizer

$$R_T \leq \psi_T(\mathbf{w}^*) - \psi_1(\mathbf{w}_1) + \sum_{t=1}^T \left(L_{t-1}^\psi(\mathbf{w}_t) - L_t^\psi(\mathbf{w}_{t+1}) + \ell_t(\mathbf{w}_t) \right)$$



Analysis of FTRL with time-varying regularizer

$$R_T \leq \psi_T(\mathbf{w}^*) - \psi_1(\mathbf{w}_1) + \sum_{t=1}^T \left(L_{t-1}^\psi(\mathbf{w}_t) - L_t^\psi(\mathbf{w}_{t+1}) + \ell_t(\mathbf{w}_t) \right)$$

We now bound the terms $L_{t-1}^\psi(\mathbf{w}_t) - L_t^\psi(\mathbf{w}_{t+1}) + \ell_t(\mathbf{w}_t)$



Analysis of FTRL with time-varying regularizer (cont.)

- ▶ Assume ψ_t is μ_t -strongly convex and $\psi_t \leq \psi_{t+1}$ for $t \geq 1$



Analysis of FTRL with time-varying regularizer (cont.)

- ▶ Assume ψ_t is μ_t -strongly convex and $\psi_t \leq \psi_{t+1}$ for $t \geq 1$
- ▶ Recall $f(\mathbf{w}) - f(\mathbf{w}^*) \geq \frac{\mu}{2} \|\mathbf{w} - \mathbf{w}^*\|^2$ for f μ -strongly convex and $\mathbf{w}^* = \operatorname{argmin}_{\mathbf{w} \in \mathcal{V}} f(\mathbf{w})$



Analysis of FTRL with time-varying regularizer (cont.)

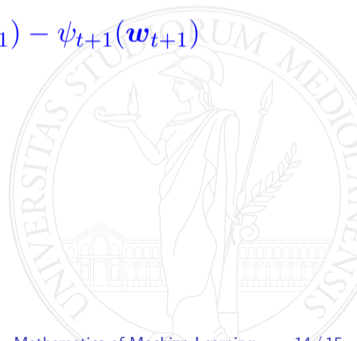
- ▶ Assume ψ_t is μ_t -strongly convex and $\psi_t \leq \psi_{t+1}$ for $t \geq 1$
- ▶ Recall $f(\mathbf{w}) - f(\mathbf{w}^*) \geq \frac{\mu}{2} \|\mathbf{w} - \mathbf{w}^*\|^2$ for f μ -strongly convex and $\mathbf{w}^* = \operatorname{argmin}_{\mathbf{w} \in \mathcal{V}} f(\mathbf{w})$
- ▶ Let $\mathbf{w}_t^* = \operatorname{argmin}_{\mathbf{w} \in \mathcal{V}} L_{t-1}^\psi(\mathbf{w}) + \ell_t(\mathbf{w})$



Analysis of FTRL with time-varying regularizer (cont.)

- ▶ Assume ψ_t is μ_t -strongly convex and $\psi_t \leq \psi_{t+1}$ for $t \geq 1$
- ▶ Recall $f(\mathbf{w}) - f(\mathbf{w}^*) \geq \frac{\mu}{2} \|\mathbf{w} - \mathbf{w}^*\|^2$ for f μ -strongly convex and $\mathbf{w}^* = \operatorname{argmin}_{\mathbf{w} \in \mathbb{V}} f(\mathbf{w})$
- ▶ Let $\mathbf{w}_t^* = \operatorname{argmin}_{\mathbf{w} \in \mathbb{V}} L_{t-1}^\psi(\mathbf{w}) + \ell_t(\mathbf{w})$

$$\begin{aligned} & L_{t-1}^\psi(\mathbf{w}_t) - L_t^\psi(\mathbf{w}_{t+1}) + \ell_t(\mathbf{w}_t) \\ &= (L_{t-1}^\psi(\mathbf{w}_t) + \ell_t(\mathbf{w}_t)) - (L_{t-1}^\psi(\mathbf{w}_{t+1}) + \ell_t(\mathbf{w}_{t+1})) + \psi_t(\mathbf{w}_{t+1}) - \psi_{t+1}(\mathbf{w}_{t+1}) \end{aligned}$$



Analysis of FTRL with time-varying regularizer (cont.)

- ▶ Assume ψ_t is μ_t -strongly convex and $\psi_t \leq \psi_{t+1}$ for $t \geq 1$
- ▶ Recall $f(\mathbf{w}) - f(\mathbf{w}^*) \geq \frac{\mu}{2} \|\mathbf{w} - \mathbf{w}^*\|^2$ for f μ -strongly convex and $\mathbf{w}^* = \operatorname{argmin}_{\mathbf{w} \in \mathbb{V}} f(\mathbf{w})$
- ▶ Let $\mathbf{w}_t^* = \operatorname{argmin}_{\mathbf{w} \in \mathbb{V}} L_{t-1}^\psi(\mathbf{w}) + \ell_t(\mathbf{w})$

$$\begin{aligned} & L_{t-1}^\psi(\mathbf{w}_t) - L_t^\psi(\mathbf{w}_{t+1}) + \ell_t(\mathbf{w}_t) \\ &= (L_{t-1}^\psi(\mathbf{w}_t) + \ell_t(\mathbf{w}_t)) - (L_{t-1}^\psi(\mathbf{w}_{t+1}) + \ell_t(\mathbf{w}_{t+1})) + \psi_t(\mathbf{w}_{t+1}) - \psi_{t+1}(\mathbf{w}_{t+1}) \\ &\leq (L_{t-1}^\psi(\mathbf{w}_t) + \ell_t(\mathbf{w}_t)) - (L_{t-1}^\psi(\mathbf{w}_t^*) + \ell_t(\mathbf{w}_t^*)) \quad (\text{minimality of } \mathbf{w}_t^* \text{ and condition on } \psi_t) \end{aligned}$$

Analysis of FTRL with time-varying regularizer (cont.)

- ▶ Assume ψ_t is μ_t -strongly convex and $\psi_t \leq \psi_{t+1}$ for $t \geq 1$
- ▶ Recall $f(\mathbf{w}) - f(\mathbf{w}^*) \geq \frac{\mu}{2} \|\mathbf{w} - \mathbf{w}^*\|^2$ for f μ -strongly convex and $\mathbf{w}^* = \operatorname{argmin}_{\mathbf{w} \in \mathbb{V}} f(\mathbf{w})$
- ▶ Let $\mathbf{w}_t^* = \operatorname{argmin}_{\mathbf{w} \in \mathbb{V}} L_{t-1}^\psi(\mathbf{w}) + \ell_t(\mathbf{w})$

$$\begin{aligned} & L_{t-1}^\psi(\mathbf{w}_t) - L_t^\psi(\mathbf{w}_{t+1}) + \ell_t(\mathbf{w}_t) \\ &= (L_{t-1}^\psi(\mathbf{w}_t) + \ell_t(\mathbf{w}_t)) - (L_{t-1}^\psi(\mathbf{w}_{t+1}) + \ell_t(\mathbf{w}_{t+1})) + \psi_t(\mathbf{w}_{t+1}) - \psi_{t+1}(\mathbf{w}_{t+1}) \\ &\leq (L_{t-1}^\psi(\mathbf{w}_t) + \ell_t(\mathbf{w}_t)) - (L_{t-1}^\psi(\mathbf{w}_t^*) + \ell_t(\mathbf{w}_t^*)) \quad (\text{minimality of } \mathbf{w}_t^* \text{ and condition on } \psi_t) \\ &(L_{t-1}^\psi(\mathbf{w}_t) + \ell_t(\mathbf{w}_t)) - (L_{t-1}^\psi(\mathbf{w}_t^*) + \ell_t(\mathbf{w}_t^*)) \geq \frac{\mu_t}{2} \|\mathbf{w}_t - \mathbf{w}_t^*\|^2 \quad (\text{s.c.} + \text{minimality of } \mathbf{w}_t^*) \end{aligned}$$

Analysis of FTRL with time-varying regularizer (cont.)

- ▶ Assume ψ_t is μ_t -strongly convex and $\psi_t \leq \psi_{t+1}$ for $t \geq 1$
- ▶ Recall $f(\mathbf{w}) - f(\mathbf{w}^*) \geq \frac{\mu}{2} \|\mathbf{w} - \mathbf{w}^*\|^2$ for f μ -strongly convex and $\mathbf{w}^* = \operatorname{argmin}_{\mathbf{w} \in \mathbb{V}} f(\mathbf{w})$
- ▶ Let $\mathbf{w}_t^* = \operatorname{argmin}_{\mathbf{w} \in \mathbb{V}} L_{t-1}^\psi(\mathbf{w}) + \ell_t(\mathbf{w})$

$$\begin{aligned} & L_{t-1}^\psi(\mathbf{w}_t) - L_{t-1}^\psi(\mathbf{w}_{t+1}) + \ell_t(\mathbf{w}_t) \\ &= (L_{t-1}^\psi(\mathbf{w}_t) + \ell_t(\mathbf{w}_t)) - (L_{t-1}^\psi(\mathbf{w}_{t+1}) + \ell_t(\mathbf{w}_{t+1})) + \psi_t(\mathbf{w}_{t+1}) - \psi_{t+1}(\mathbf{w}_{t+1}) \\ &\leq (L_{t-1}^\psi(\mathbf{w}_t) + \ell_t(\mathbf{w}_t)) - (L_{t-1}^\psi(\mathbf{w}_t^*) + \ell_t(\mathbf{w}_t^*)) \quad (\text{minimality of } \mathbf{w}_t^* \text{ and condition on } \psi_t) \\ &(L_{t-1}^\psi(\mathbf{w}_t) + \ell_t(\mathbf{w}_t)) - (L_{t-1}^\psi(\mathbf{w}_t^*) + \ell_t(\mathbf{w}_t^*)) \geq \frac{\mu_t}{2} \|\mathbf{w}_t - \mathbf{w}_t^*\|^2 \quad (\text{s.c.} + \text{minimality of } \mathbf{w}_t^*) \\ &L_{t-1}^\psi(\mathbf{w}_t) - L_{t-1}^\psi(\mathbf{w}_t^*) + \ell_t(\mathbf{w}_t) - \ell_t(\mathbf{w}_t^*) \leq G \|\mathbf{w}_t - \mathbf{w}_t^*\| \quad (\text{minimality of } \mathbf{w}_t + \text{Lip.}) \end{aligned}$$

Analysis of FTRL with time-varying regularizer (cont.)

- ▶ Assume ψ_t is μ_t -strongly convex and $\psi_t \leq \psi_{t+1}$ for $t \geq 1$
- ▶ Recall $f(\mathbf{w}) - f(\mathbf{w}^*) \geq \frac{\mu}{2} \|\mathbf{w} - \mathbf{w}^*\|^2$ for f μ -strongly convex and $\mathbf{w}^* = \operatorname{argmin}_{\mathbf{w} \in \mathbb{V}} f(\mathbf{w})$
- ▶ Let $\mathbf{w}_t^* = \operatorname{argmin}_{\mathbf{w} \in \mathbb{V}} L_{t-1}^\psi(\mathbf{w}) + \ell_t(\mathbf{w})$

$$\begin{aligned} & L_{t-1}^\psi(\mathbf{w}_t) - L_t^\psi(\mathbf{w}_{t+1}) + \ell_t(\mathbf{w}_t) \\ &= (L_{t-1}^\psi(\mathbf{w}_t) + \ell_t(\mathbf{w}_t)) - (L_{t-1}^\psi(\mathbf{w}_{t+1}) + \ell_t(\mathbf{w}_{t+1})) + \psi_t(\mathbf{w}_{t+1}) - \psi_{t+1}(\mathbf{w}_{t+1}) \\ &\leq (L_{t-1}^\psi(\mathbf{w}_t) + \ell_t(\mathbf{w}_t)) - (L_{t-1}^\psi(\mathbf{w}_t^*) + \ell_t(\mathbf{w}_t^*)) \quad (\text{minimality of } \mathbf{w}_t^* \text{ and condition on } \psi_t) \\ &(L_{t-1}^\psi(\mathbf{w}_t) + \ell_t(\mathbf{w}_t)) - (L_{t-1}^\psi(\mathbf{w}_t^*) + \ell_t(\mathbf{w}_t^*)) \geq \frac{\mu_t}{2} \|\mathbf{w}_t - \mathbf{w}_t^*\|^2 \quad (\text{s.c. + minimality of } \mathbf{w}_t^*) \\ &L_{t-1}^\psi(\mathbf{w}_t) - L_{t-1}^\psi(\mathbf{w}_t^*) + \ell_t(\mathbf{w}_t) - \ell_t(\mathbf{w}_t^*) \leq G \|\mathbf{w}_t - \mathbf{w}_t^*\| \quad (\text{minimality of } \mathbf{w}_t + \text{Lip.}) \\ &L_{t-1}^\psi(\mathbf{w}_t) - L_t^\psi(\mathbf{w}_{t+1}) + \ell_t(\mathbf{w}_t) \leq \frac{2G^2}{\mu_t} \end{aligned}$$

Regret bound

Assume $\psi \geq 0$ is μ -strongly convex and $\psi_t = \frac{\psi}{\eta_{t-1}}$ where $\eta_t \leq \eta_{t-1}$ for $t \geq 1$



Regret bound

Assume $\psi \geq 0$ is μ -strongly convex and $\psi_t = \frac{\psi}{\eta_{t-1}}$ where $\eta_t \leq \eta_{t-1}$ for $t \geq 1$

$$R_T \leq \psi_T(\mathbf{w}^*) + 2G^2 \sum_{t=1}^T \frac{1}{\mu_t}$$

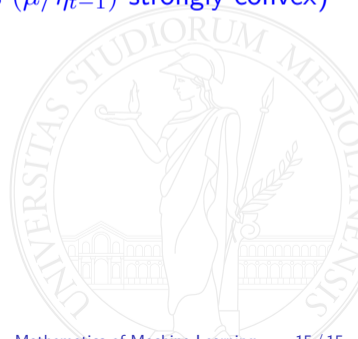


Regret bound

Assume $\psi \geq 0$ is μ -strongly convex and $\psi_t = \frac{\psi}{\eta_{t-1}}$ where $\eta_t \leq \eta_{t-1}$ for $t \geq 1$

$$\begin{aligned} R_T &\leq \psi_T(\mathbf{w}^*) + 2G^2 \sum_{t=1}^T \frac{1}{\mu_t} \\ &= \frac{D^2}{\eta_T} + 2G^2 \sum_{t=1}^T \frac{\eta_{t-1}}{\mu} \end{aligned}$$

(ψ_t is (μ/η_{t-1}) -strongly convex)



Regret bound

Assume $\psi \geq 0$ is μ -strongly convex and $\psi_t = \frac{\psi}{\eta_{t-1}}$ where $\eta_t \leq \eta_{t-1}$ for $t \geq 1$

$$\begin{aligned} R_T &\leq \psi_T(\mathbf{w}^*) + 2G^2 \sum_{t=1}^T \frac{1}{\mu_t} \\ &= \frac{D^2}{\eta_T} + 2G^2 \sum_{t=1}^T \frac{\eta_{t-1}}{\mu} \\ &= GD \sqrt{\frac{T}{\mu}} + 2GD \sqrt{\frac{1}{\mu}} \sum_{t=1}^T \sqrt{t} \end{aligned}$$

(ψ_t is (μ/η_{t-1}) -strongly convex)

$$\left(\eta_{t-1} = \frac{D}{G} \sqrt{\frac{\mu}{t}}\right)$$

Regret bound

Assume $\psi \geq 0$ is μ -strongly convex and $\psi_t = \frac{\psi}{\eta_{t-1}}$ where $\eta_t \leq \eta_{t-1}$ for $t \geq 1$

$$\begin{aligned} R_T &\leq \psi_T(\mathbf{w}^*) + 2G^2 \sum_{t=1}^T \frac{1}{\mu_t} \\ &= \frac{D^2}{\eta_T} + 2G^2 \sum_{t=1}^T \frac{\eta_{t-1}}{\mu} \\ &= GD \sqrt{\frac{T}{\mu}} + 2GD \sqrt{\frac{1}{\mu}} \sum_{t=1}^T \sqrt{t} \\ &\leq 5GD \sqrt{\frac{T}{\mu}} \end{aligned}$$

(ψ_t is (μ/η_{t-1}) -strongly convex)

using $\sum_{t=1}^T \sqrt{t} \leq 2\sqrt{T}$

